

PAPER • OPEN ACCESS

Smoothing parameters of penalized spline nonparametric regression model using linear mixed model

To cite this article: D R S Saputro *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **243** 012040

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the **collection** - download the first chapter of every title for free.

Smoothing parameters of penalized spline nonparametric regression model using linear mixed model

D R S Saputro^{1*}, H Lukmawati², P Widyaningsih³

^{1,2,3} Universitas Sebelas Maret, Solo, Indonesia

E-mail: dewiretnoss@staff.uns.ac.id

Abstract. Nonparametric regression is method of regression approach for data patterns of which regression curve is unknown. Meanwhile, spline is a type of segmented polynomial piecewise. Such segmented nature gives more flexibility than common polynomial models so that spline can adapt more effectively to local characteristics of function or data. The use of spline focuses on presence of behavior and patterns of data which in certain area has different characteristics from other areas. Data matching can be done by observing points on data in which an extreme change occurs in an area and consequently data patterns in each area differ. One approach using nonparametric regression is spline regression with penalized spline (P-spline). It is defined as a regression determined using method of least squares and roughness penalty. The article examines smoothing parameters of penalized spline nonparametric regression model. P-spline can be represented in linear mixed model with components of variance to control nonlinearity level of estimators of smooth functions. P-spline using linear mixed model approach can be estimated using either maximum likelihood or residual maximum likelihood. The results of research show that smooth function of best linear unbiased predictor in linear mixed model is equivalent to estimators of penalized spline regression.

1. Introduction

Regression model is used to model the correlation between response variable and one or more explanatory variables. Such method was developed from parametric regression model to nonparametric regression model. In the parametric regression model, regression curve estimation is equivalent to estimation of parameters in the model ([1]). In order to avoid the use of strict assumptions, nonparametric regression approach is an alternative that can be used. Nonparametric regression refers to a modeling method which is not bound to assumptions of certain regression equation providing high flexibility in estimating a model. Several nonparametric regression estimation methods which can be used include the Fourier series, Spline and Kernel ([2]). Spline regression is a piecewise polynomial regression analysis method, polynomial pieces with segmented nature on an interval formed on knots. Knots are join points which emerge due to the change in behavior patterns at different intervals. Spline is superior in dealing with data patterns indicating either sharp raise or drop with the help of knots, and the resulted curve is relatively smooth ([3]). Spline estimators tend to seek their estimation by themselves wherever the data move to obtain a model which is appropriate with the forms of data. A criterion which can be used in the selection of optimal knots is Generalized Cross Validation (GCV) ([1]). Spline refers to the segmented or smoothly-joined piecewise polynomial



model which can produce regression functions fit to the data. To estimate spline depends on the knots, the join points which emerge due to the change in behavior patterns of a function at different intervals. In contrast to in parametric regression approach, the shape of regression curve in nonparametric regression is assumed to be unknown. Nonparametric regression curve is only assumed to be smooth in the sense that it is contained in certain function space. Data are expected to seek their own forms of estimation without being influenced by the subjectivity factor in research design. Therefore, nonparametric regression approach has high flexibility ([2]). Spline nonparametric regression is a segmented regression giving discretion to different polynomial functions on each segment with smoothing splines to result in curve along observation x .

An approach using nonparametric regression is spline regression with penalized spline method. Spline regression refers to an approach towards data plot by taking smoothness of curve into account. Spline is a segmented polynomial model in which the characteristics of the segments contribute to better flexibility than common polynomial models. According to [1, 2], spline in nonparametric regression has high flexibility and an ability to estimate the behavior of data which tends to be different at different intervals.

There are two types of nonparametric model approach using spline: spline regression and smoothing splines. Reference [4], spline regression has excellent statistical and visual interpretations. Moreover, the spline regression is able to handle the characteristics of data and smooth functions and has less capability to handle data of which behavior changes at certain subinterval. It requires a relatively small number of knots and can be estimated using the method of least squares, while spline smoother needs a large number of knots and the smoothness of curve is determined by smoothing parameters and penalized function. Reference [5] combined the two spline approaches to be P-spline. Similarly, [6] used the term penalized spline, or P-spline. P-spline combines two advantages: parametric estimation on spline regression and flexible adaptation towards the smoothness level of curve resulted from roughness penalty on spline smoother. The penalized spline regression has a simple mathematical correlation with the linear mixed models, and therefore the smoothing parameters are directly connected to the variance components and the linear mixed models. In the article, we show that smoothing functions for Best Linear Unbiased Predictor (BLUP) in linear mixed models are equivalent to estimators of penalized spline regression.

2. Theoretical review

2.1. *Spline nonparametric regression*

Reference [3] points out that the nonparametric approach to estimate regression curves is intended to provide a good model to find out the correlation between two variables and to give observation model to result in appropriate prediction. Outlines that nonparametric regression is an approach for the data patterns of which regression curve is unknown and the absence of information on distribution of data patterns [2].

The idea in nonparametric regression approach is the smoothness of curve. It means that through nonparametric regression approach it is assumed that regression curve resulted from the data patterns is smooth due to certain function, and consequently the nonparametric regression has flexibility. The smoothness in nonparametric regression method aims at lowering variance in data and estimating behavior of data which tends to be different and have an influence, so the characteristics of data can be clearly seen.

2.2. *Optimal knots*

The best spline estimator is obtained by using optimal knots. Knots exist in the change in behavior patterns of a function. According to [6], to select optimal knots, generalized cross validation (GCV) method is used. The formula of GCV is expressed:

$$GCV(K_1, K_2, \dots, K_l) = \frac{MSE(K_1, K_2, \dots, K_l)}{\left(\frac{1}{n} \text{tr}[I - A(K_1, K_2, \dots, K_l)]\right)^{2l}}$$

Where $K_1, K_2, \dots, K_l$ represent knots and matrix $A(K_1, K_2, \dots, K_l)$ is obtained from the formula $X_K(X_K^T X_K)^{-1} X_K^T$ and mean squared error (MSE) is the average of the squares of the errors or deviations. The formula of MSE is:

$$MSE(K_1, K_2, \dots, K_l) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

3. Research Method

This research is a research of theoretical study in showing that smoothing functions for Best Linear Unbiased Predictor (BLUP) in linear mixed models are equivalent to estimators of penalized spline regression. The basic material on which this research is the scientific work of several experts presented and published in journals, bulletins and books. The research method used is literature study from several literatures, by studying the scientific papers that have been collected. The research step is to derive estimation of parameter of penalized spline regression and linear mixed model, then to determine correlation between penalized spline regression and linear mixed model method. Finally, we examined the connection between smoothing splines and a linear mixed effects model, and conducted a study on best linear unbiased predictor (BLUP) from linear mixed model.

4. Result and Discussion

4.1. Penalized spline regression

Given that (x_i, y_i) is measure on explanatory variable x and response variable y for $1 \leq i \leq n$, the functional correlation between x and y is modeled as:

$$y_i = s(x_i) + \varepsilon_i \quad (1)$$

where s represents smooth function. Equation (1) is the most simple form of nonparametric regression. Given that smooth function is estimated using spline regression model:

$$s(x; \beta) = \beta_0 + \beta_1 X + \dots + \beta_p X^p + \sum_{k=1}^p u_{pk} (x - x_k)_+^p \quad (2)$$

where $\beta = (\beta_0, \dots, \beta_p, u_{p1}, \dots, u_{pk})$ represents coefficient vector of spline regression, $p \geq 1$ are positive integers, $(w)_+^p = w^p I(w \geq 0)$ is p -degree segmented basis function.

Parameter estimator $\hat{\beta}$ is determined by minimizing the penalized sum of squares. According to [7], $J(s)$ is defined as:

$$J(s) = \sum_{i=1}^n (y_i - s(x_i; \beta))^2 + \lambda \beta' \mathcal{D} \beta \quad (3)$$

where λ is a smoothing parameter, and $\mathcal{D} = \text{diag}(\mathcal{O}_{p+1}, I_k)$. The first term $J(s)$ represents the sum of squares error and the second term is the roughness penalty. The determination on equation (3) is the combination of spline regression model and smoothing spline model.

The smoothing parameter $\lambda \geq 0$ depicts the rate of exchange between the sum of squares error and the local variance. If λ has a large value, then the main component $J(s)$ is the roughness penalty, and therefore curve will look smooth. In contrast, if λ has a small value, then the main component $J(s)$ is the component of the sum of squares error, and therefore curve will seem rough. Given that T is the design matrix for spline regression, where the i^{th} term of matrix T is:

$$T = \mathbf{1}, x, \dots, x^p, (x - K_1)_+^p, \dots, (x - K_K)_+^p$$

in matrix notation, $J(s)$ is expressed:

$$\|y - T\beta\|^2 + \lambda\beta'D\beta. \quad (4)$$

To minimize $J(\beta)$, it is derived towards β with principle $\frac{\partial J}{\partial \beta} = 0$ and therefore the following equation is obtained:

$$\frac{\partial J}{\partial \beta} = -2T'y + 2T'T\beta + 2\lambda\beta'D\beta = 0$$

$$T'T\beta + \lambda\beta'D\beta = T'y$$

$$(T'T + \lambda\beta'D)\beta = T'y$$

Minimizing equation (4) results in an estimator for parameter $\hat{\beta}$:

$$\hat{\beta}(T'T + \lambda D)^{-1}T'y$$

Thus, the estimator of penalized spline regression is obtained.

$$\hat{y} = T\hat{\beta}(T'T + \lambda D)^{-1}T'y \quad (5)$$

4.2. Linear mixed model

The linear mixed model is generally defined as:

$$y = X\beta + Zu + \varepsilon, \text{ where } \begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \quad (6)$$

where X represents the design matrix of the observed effects, β is the vector of unknown fixed effect parameters, Z is the design matrix of observed random effects, u depicts the vector of unknown random effects, and is the vector of unknown errors. Thus, the median and the matrix for y are $E(y) = X\beta$ and $Cov(y) = V = ZGZ' + R$

For nonsingular G , the joint density of y and u is:

$$\begin{aligned} f_{y,u} &= f_{y|u} \cdot f_u \\ &= |2\pi R|^{-1/2} \\ &\exp\left(-\frac{1}{2}(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu)\right) \\ &\times |2\pi G|^{-1/2} \exp\left(-\frac{1}{2}u'G^{-1}u\right) \end{aligned}$$

The logarithm of the joint density function is expressed:

$$\begin{aligned} \ln(f_{y,u}) &= -\frac{1}{2} \ln |2\pi R| \\ &- \frac{1}{2} (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) \\ &- \frac{1}{2} \ln |2\pi G| - \frac{1}{2} u'G^{-1}u \end{aligned}$$

The partial derivative of $\ln(f_{y,u})$ towards u is:

$$\frac{\partial f_{y,u}}{\partial u} = Z'R^{-1}(y - X\beta - Zu) + G^{-1}u$$

and the partial derivative of $\ln(f_{y,u})$ towards β is:

$$\frac{\partial f_{y,u}}{\partial \beta} = X'R^{-1}(y - X\beta - Zu) \quad (7)$$

To determine values of u and β , maximizing equation (7) results in the equation:

$$Z'R^{-1}y - Z'R^{-1}X\beta - (Z'R^{-1}Z + G^{-1})u = 0 \quad (8)$$

$$X'R^{-1}y - X'R^{-1}X\beta - Z'R^{-1}Zu = 0 \quad (9)$$

In the form of matrix, (8) and (9) are expressed:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad (10)$$

If matrix G and R were known, then the estimator for parameters β and u is:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X' \\ Z' \end{bmatrix} R^{-1}y$$

Equation (9) is derived to obtain u , and therefore the equation (10) is solved below:

$$u = (Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}(y - X\beta) \quad (11)$$

Next, equation (11) is substituted into equation (8).

$$X'R^{-1}y - X'R^{-1}X\beta - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}(y - X\beta) = 0$$

The following equation is obtained:

$$X'(R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1})y = X'(R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1})X\beta \quad (12)$$

Where $V = ZGZ' + R$. From equation (12), the following equation is obtained:

$$R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1} = V^{-1}$$

To derive

$$X'V^{-1}y = X'V^{-1}X\beta \quad (13)$$

From equation (13), the following equation is resulted:

$$\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}y \quad (14)$$

Afterwards, equation (13) and equation (11) are paired and using some simple algebra the following equation is obtained:

$$\begin{aligned} \hat{u} &= (Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}(y - X\hat{\beta}) = (Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}VV^{-1}(y - X\hat{\beta}) \\ &= (Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}(ZGZ' + R)V^{-1}(y - X\hat{\beta}) \\ &= (Z'R^{-1}Z + G^{-1})^{-1} (Z'R^{-1}ZGZ' + Z')V^{-1}(y - X\hat{\beta}) \\ &= (Z'R^{-1}Z + G^{-1})^{-1} (Z'R^{-1}Z + G^{-1})GZV^{-1}(y - X\hat{\beta}) \\ \hat{u} &= GZV^{-1}(y - X\hat{\beta}) \end{aligned}$$

Best linear unbiased predictor (BLUP) for β is identical to the solution of GLS, and BLUP for u is BLUP (u) with substitution $\hat{\beta}$ and β . One of the most simple ways to obtain BLUP is by applying Henderson's justification with distribution assumptions ([8]), i.e. $y|u \sim N(X\beta + Zu, R)$ and $u \sim N(0, G)$.

Maximizing function (y, u) on unknown β and u results in:

$$(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u \quad (15)$$

Equation (15) indicates that BLUP estimation of (β, u) comprises GLS and a penalty term. BLUP of (β, u) is expressed:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = (C'R^{-1} + B)^{-1} C'R^{-1}y$$

Where $C = [X \ Z]$ and $B = \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} \end{bmatrix}$ and therefore BLUP

$$(y) = X\hat{\beta} + Z\hat{u} = C(C'R^{-1}C)^{-1} C'R^{-1}y$$

The most commonly used methods in the estimation of variance-covariance matrix on linear mixed method are ML and REML methods. REML method generates unbiased estimators for parameters of the variance-covariance matrix, while ML results in biased estimators. The estimators of ML from are determined according to distribution model $y \sim N(X\beta + Zu, V)$

The log-likelihood function of y according to the distribution model is:

$$\ell_{ML}(\beta, V) = -\frac{1}{2} \{ \log|V| + (y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) + n \log(2\pi) \} \quad (16)$$

ML estimators for $(\beta, V)'$ are obtained by maximizing $\ell_{ML}(\beta, V)$. Optimization on β is done given that V is fixed, and results in $\ell_{ML}(\beta, V)$ maximum $\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}y$, similar to BLUP on equation (13).

The substitution of $\hat{\beta}$ on equation (16) generates log-likelihood for V :

$$\begin{aligned}\ell_p(V) &= -\frac{1}{2}\{\log|V| + (y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) + n\log(2\pi)\} \\ &= -\frac{1}{2}\{\log|V| + y'V^{-1}(1 - X[X'V^{-1}X]^{-1}X'V^{-1})y\} - n\log(2\pi)\end{aligned}\quad (17)$$

ML estimators for parameters in can be obtained by maximizing equation (17) towards its parameters. The reduction in REML by maximizing log-likelihood function of linear combination of element which does not depend on . The function of REML is:

$$\begin{aligned}\ell_{REML}(V) &= -\frac{1}{2}\log|V| + \log|X'V^{-1}X| + (y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) + (n - p)\log(2\pi) \\ &= \ell_p(V) - \frac{1}{2}\log|X'V^{-1}X| + \frac{p}{2}\log(2\pi)\end{aligned}$$

The main benefit of REML compared to ML is that REML considers degrees of freedom of fixed effects in linear mixed model.

4.3. The correlation between penalized spline regression and linear mixed model method

The correlation between spline regression and linear mixed model has been discussed by [9], who examined the connection between smoothing splines and a linear mixed effects model, and [10], who conducted a study on best linear unbiased predictor (BLUP) from linear mixed model. Reference [11] extended the mixed model setup. The connection between penalized spline regression and linear mixed model includes penalty coefficient on equation equivalent to treating the coefficient as random effects on linear mixed model on equation (11). Given that parameter $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, $u = (u_{p1}, \dots, u_{pk})$ are defined, and design matrix are:

$$X = \begin{bmatrix} 1 & t_1 & \dots & t_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \dots & t_n^p \end{bmatrix} \text{ and } Z = \begin{bmatrix} (t_1 - k_1)_+^p & \dots & (t_1 - k_k)_+^p \\ \vdots & \ddots & \vdots \\ (t_n - k_1)_+^p & \dots & (t_n - k_k)_+^p \end{bmatrix}$$

The criterion of penalized spline on equation (9) if divided by σ_ε^2 can be expressed:

$$\frac{1}{\sigma_\varepsilon^2} \|y - X\beta - Zu\|^2 + \frac{\lambda_\varepsilon}{\sigma_\varepsilon^2} \|u\|^2 \quad (18)$$

Equation (18) is similar to BLUP criterion of linear mixed model on equation (14) by treating u as the coefficient of random effects with $(u) = \sigma_\varepsilon^2 I$, where $\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{\lambda^2}$, and therefore the representation of spline regression in the form of linear mixed model is:

$$y = X\beta + Zu + \varepsilon, \text{ where } Cov \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix}$$

BLUP for function $s(x) = (s(x_1), \dots, s(x_k))'$ is given by:

$$\hat{y} = X\hat{\beta} + Z\hat{u} \quad (19)$$

where $\hat{\beta} = (X'(\sigma_\varepsilon^2 ZZ' + \sigma_\varepsilon^2 I)^{-1}X)^{-1}X'(\sigma_\varepsilon^2 ZZ' + \sigma_\varepsilon^2 I)^{-1}y$ and $\hat{u} = \sigma_\varepsilon^2 Z(\sigma_\varepsilon^2 ZZ' + \sigma_\varepsilon^2 I)^{-1}(y - X\hat{\beta})$

Solution $\hat{y}$ on equation (18) is expressed:

$$\hat{y} = C(C'C + \lambda D)^{-1}Cy \quad (20)$$

Where $C = [X \ Z]$, $D = \text{diag}(\mathcal{O}_{p+1}, I_k)$ and $\lambda^2 = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2}$

5. Conclusion

To conclude from the results and discussion, equation (20) and equation (5) indicate that equation (20) is equivalent to solution of penalized spline regression on equation (5). Such evidence reveals that smoothing function for BLUP on linear mixed model is equivalent to estimators of penalized spline regression.

Acknowledgement

In the research, financing and writing of this paper, the authors would like to thank the two institutions : the ministry of research, technology and higher education of the Republic of Indonesia; research institute and community services of Universitas Sebelas Maret through superior basic research grants from universities with contract number 474 / UN27.21 /PP /2018.

References

- [1] Budiantara I N 2006 *Regresi nonparametrik dalam statistika* (Makassar: FMIPA UNM Makassar)
- [2] Eubank R L 1998 *Spline smoothing and nonparametric regression* (AS: Marcel Dekker New York)
- [3] Hardle W 1994 *Applied nonparametric regression* (New York: Cambridge University Press New York)
- [4] Cox D D & O'Sullivan F 1996 *Journal of Multivariate Analysis* **Vol.56** 185-206
- [5] Eilers PHC & Marx BD 1996 *Statistical Science* **Vol.11** 89-121
- [6] Wahba G 1990 *Spline models for observational data* (SIAM Pennsylvania)
- [7] Ruppert D, Carroll RJ 1997 Penalized regression splines *Unpublished manuscript*
- [8] Speed T 1991 *Statistical Science* **Vol.6** 42-44
- [9] Vebyla A P, Cullis B R, Kenward M G, Welham S J 1999 *Journal of The Royal Statistics Society* **Vol.C 48** 269-312
- [10] Greek P J, Silverman B W 1994 *Nonparametric regression and generalized linear models : A roughness penalty approach* (London: Chapman & Hall London)
- [11] Wang Y 1998 *Journal of the Royal Statistical Society* **Vol.B. 60** 159-174